

---

# Diseño e implementación de un repositorio de acceso público de datos y señales relacionados al estudio de la epilepsia

---

Samuel Josué Silvestre Mendoza





UNIVERSIDAD DEL VALLE DE GUATEMALA  
Facultad de Ingeniería



**Diseño e implementación de un repositorio de acceso público  
de datos y señales relacionados al estudio de la epilepsia**

Trabajo de graduación presentado por Samuel Josué Silvestre Mendoza  
para optar al grado académico de Licenciado en Ingeniería Mecatronica

Guatemala,

2022



Vo.Bo.:

(f) \_\_\_\_\_  
Ing. Luis Alberto Rivera Estrada

Tribunal Examinador:

(f) \_\_\_\_\_  
Ing. Luis Alberto Rivera Estrada

(f) \_\_\_\_\_

(f) \_\_\_\_\_

Fecha de aprobación: Guatemala, de de 2022.



<b>Lista de figuras</b>	<b>VII</b>
<b>Resumen</b>	<b>IX</b>
<b>Abstract</b>	<b>XI</b>
<b>1. Introducción</b>	<b>1</b>
<b>2. Antecedentes</b>	<b>3</b>
<b>3. Justificación</b>	<b>7</b>
<b>4. Objetivos</b>	<b>9</b>
<b>5. Alcance</b>	<b>11</b>
<b>6. Marco teórico</b>	<b>13</b>
6.1. Epilepsia . . . . .	13
6.1.1. Tipos de epilepsia . . . . .	13
6.2. Electroencefalograma . . . . .	14
6.3. Base de datos . . . . .	15
6.3.1. Conceptos básicos . . . . .	15
6.4. Base de datos en la nube . . . . .	15
6.4.1. Ventajas . . . . .	16
6.5. Servicios . . . . .	16
6.5.1. Dspace . . . . .	16
6.6. Dataverse . . . . .	17
6.6.1. DataLab . . . . .	17
6.7. Aprendizaje Automático y la Epilepsia . . . . .	17
<b>7. Análisis Servicios:</b>	<b>19</b>
7.1. Amazon AWS . . . . .	19
7.2. Microsoft Azure . . . . .	20
7.3. Oracle . . . . .	20

7.4. introducción a DataLab . . . . .	20
<b>8. Prototipo 1</b>	<b>21</b>
<b>9. Prototipo 2: Dataverse</b>	<b>23</b>
9.1. Dataverse: Pruebas . . . . .	24
9.2. Dataverse: Análisis . . . . .	25
9.3. Dataverse: Usuario . . . . .	26
<b>10. Prototipo 3: Nuevo Dataverse</b>	<b>29</b>
<b>11. Nuevo dataverse de Pruebas</b>	<b>31</b>
<b>12. Cómo Utilizar el Dataverse</b>	<b>33</b>
<b>13. Conclusiones</b>	<b>37</b>
<b>14. Recomendaciones</b>	<b>39</b>
<b>15. Bibliografía</b>	<b>41</b>
<b>16. Anexos</b>	<b>43</b>
16.1. Repositorio Dataverse . . . . .	43
16.2. Repositorio Github . . . . .	43



---

## Lista de figuras

---

1.	Modelo relacional utilizado en la fase anterior para la base de datos [7]. . . . .	5
2.	Actividad cerebral registrada por electroencefalograma [10]. . . . .	14
3.	Servicios de Milab [9]. . . . .	18
4.	tabla para pruebas. . . . .	22
5.	app ddesarrollada en visual studio. . . . .	22
6.	Relación entre los <i>Dataverse</i> del repositorio. . . . .	23
7.	Ejemplo dataset de edf de un paciente. . . . .	25
8.	Ejemplo dataset Analisis de un edf. . . . .	26
9.	Ejemplo perfil de usuario. . . . .	27
10.	Ejemplo Nuevo formato de Dataset de Análisis . . . . .	30
11.	Ejemplo nuevo Dataset de Pruebas . . . . .	32
12.	Página donde se crea una cuenta. . . . .	34
13.	Página donde se crea una cuenta. . . . .	34
14.	Ejemplo de como se edita un dataset. . . . .	35
15.	Archivo pdf subido al repositorio . . . . .	35
16.	Archivo rar subido al repositorio . . . . .	36



En este proyecto de realizar un repositorio para datos de investigación sobre la epilepsia, se empezó continuando lo desarrollado en la fase II por Jorge Diego Manrique Sáenz, en la tesis “Herramienta de Software con una Base de Datos Integrada para el Estudio de la Epilepsia - Fase II”, lo que se ha buscado es llevar el concepto de la base de datos a un repositorio que se encuentre en la nube, y que sea de carácter público, de manera que contribuya a la comunidad científica en el estudio de la epilepsia y sus efectos.

A partir de la herramienta HUMANA y demás herramientas que nos permiten extraer datos para estudiar la epilepsia con métodos como electroencefalogramas, se buscó preparar el mejor espacio posible para que los usuarios puedan proveer o extraer datos de manera sencilla, y puedan contribuir a la comunidad de la manera que deseen.

Se investigaron todas las opciones disponibles de desarrollo en el mercado, la seguridad que estas brindaban, el coste de ellas, y de probar estructuras de forma local. Finalmente, se escogió la herramienta “DataLab” un servicio proveído por RedCLARA , en la plataforma “Dataverse” de la universidad de Harvard. A través de su sistema de archivado, se puede ofrecer un repositorio de fácil acceso para todo tipo de archivos relacionados a la epilepsia.



Going through with the development of the face 2 , done by Jorge Diego Manrique Sáenz, on the thesis “Herramienta de Software con una Base de Datos Integrada para el Estudio de la Epilepsia - Fase II”, we searched to take the concept of the database from a local one to a public repository stored on the cloud, with the objective of making a contribution to the scientific community.

Using the tool “HUMANA “as a starting point together with other similar tools that let us study the epilepsy , with methods like an electroencephalogram, we sought to achieve the best possible space for the users so they could provide or extract data from the repository in an easy manner, and contribute to the community the way they want.

After investigating all the possible choices for our repository, how secure they are, prices, and after doing tests in a local way, we chose the tool “DataLab” provided to us by Red-CLARA in the “Dataverse” platform, from the Harvard university. Through its archiving system, we can offer an easy access repository for any kind of epilepsy related archive



Para estudiar la epilepsia y sus efectos, se han aplicado distintos algoritmos de aprendizaje automático a señales electroencefalográficas y otro tipo de señales. Esta área está en constante desarrollo, ya que la epilepsia no es un trastorno del que se conozca todo en su totalidad. Se requiere seguir investigando y progresando, para que, en un futuro, pueda ser tratada de mejor manera.

Al desarrollar investigaciones, se requiere siempre una manera de ir organizando la información que se va obteniendo, de manera que pueda ser replicable en un futuro, además de poder utilizar información o datos del pasado ya sea de propia autoría, o de terceros. El preparar un repositorio, es de suma utilidad para lograr esto.

Existen múltiples servicios por el mercado, que permiten almacenar información en bases de datos de proveedores tales como Amazon, con el fin de poder utilizar estos datos desde cualquier parte del mundo. Ofrecen distintos servicios al cliente, tanto en adaptabilidad, como en precio, como en la capacidad de brindar acceso a terceros.

Sin embargo, los servicios de esta clase no permiten una fácil accesibilidad y un ordenamiento adecuado. Con servicios como Dataverse, lo que se logra es que se le permite al usuario de la comunidad, evitar elegir entre permitir el uso de su investigación por todo el mundo, o mantener la seguridad de su propiedad intelectual. El objetivo de estos servicios de repositorio es facilitar la colaboración entre distintas entidades en desarrollar mejores formas de tratar en este caso, la epilepsia.

En este proyecto se ha buscado crear un repositorio de señales y datos biomédicos relacionados al estudio de la epilepsia, que pueda ser accedido desde cualquier parte del mundo. Esto se ha logrado al evaluar los distintos modelos de repositorio que están disponibles, obtener datos de epilepsia de HUMANA para desarrollar el prototipo, determinar la organización de los datos, y proceder a implementar el repositorio.





La epilepsia es una enfermedad que afecta al cerebro y las señales que envía, la perturbación en esta actividad neuronal causa en un aumento repentino de la actividad eléctrica que, puede llevar a causar convulsiones. Estas se alteran según el área del cerebro donde ha ocurrido [1].

A pesar de ser una de las condiciones neurológicas más comunes, más precisamente, llegando a afectar a más de 50 millones de personas alrededor del mundo (siendo más común en niños y mayores de 60 años) no se ha logrado comprender esta condición en su totalidad, teniendo como prueba de esto que, a día de hoy, no se posee un tratamiento que funcione para todos los casos. Existen más de 30 tipos de crisis epilépticas descritas con duración de unos segundos o minutos. Se pueden dividir en dos clases principales, convulsiones focales que afectan solamente una parte del cerebro, y convulsiones generalizadas que afectan todo el cerebro [2].

Cuando se habla de una base de datos relacional, se habla de que esta almacena y provee datos que están relacionados a otros. Se basan en un modelo fácil e intuitivo para acceder y almacenar la información en tablas. Cada fila es un registro con un numero de identificación único conocido como llave y posee un atributo con información especifica al registro [3].

Hay dos clases de bases de datos en la nube, está la clase tradicional que es similar a una base de datos *in-situ* y administrada internamente, con una pequeña diferencia en el origen de la infraestructura, ya que la compañía compra un espacio virtual de un proveedor de servicios en la nube. La organización se encarga de manejar la base de datos. Luego está la base de datos como servicio (DBaaS), donde el proveedor ofrece más servicios de gestión de base de datos [4].

Actualmente, en términos de procesos médicos, ya existen ejemplos de repositorios públicos de datos que ayudan a unir la comunidad en la investigación de distintas enfermedades y sus respectivos tratamientos. Tal es el caso del sitio llamado Physionet, que cuenta con un gran catalogo de archivos de señales biológicas bien categorizadas, una gran colección de software de análisis de señales fisiológicas, además de tutoriales y material educativo [5].

El Centro de Epilepsia y Neurocirugía Funcional HUMANA, es una organización que se dedica a buscar el beneficio de sus pacientes que padecen problemas Neurológicos de difícil control tales como Epilepsia, Parkinson, Tumores Cerebrales, Columna Vertebral, Movimientos Anormales entre otros. Formada por profesionales de Neurociencias, HUMANA es el Centro de Referencia en Neurociencias para Guatemala y Centro América que posee los mejores recursos para el diagnóstico y tratamiento de enfermedades cerebrales [6].

Actualmente la Universidad del Valle de Guatemala colabora con HUMANA en una estrecha relación que ha logrado que se consigan múltiples avances en áreas como la implementación de ingeniería biomédica en soluciones médicas.

El proyecto que está siendo desarrollado, y el cual continúa este trabajo, es el de desarrollar un repositorio para datos de investigación sobre la epilepsia. La tesis desarrollada por Jorge Diego Manrique [7] consistieron en mejorar lo realizado en la etapa uno desarrollada por María Fernanda Pineda [8] tratándose del desarrollo de una herramienta para análisis de señales electroencefalográficas y diseño de una base de datos para almacenar las señales grabadas por HUMANA. En la fase uno desarrollada por María Fernanda Pineda, la base de datos por medio de 35 canales, con 20 obligatorios, almacenaba la información en tres tablas, una de datos del paciente, otra de datos descriptivos de la prueba, y otra con los datos de las señales en sus diferentes canales.

En la fase dos, esto fue expandido a poder almacenar características, clasificadores, anotaciones de las señales EEG, datos confidenciales, configuración para funcionalidad “Recuperar contraseña” y accesos de los usuarios. Además de esto, se realizaron múltiples mejoras a la seguridad, manejo de usuarios e interfaz de la aplicación en Matlab utilizada para obtener y guardar los datos obtenidos con HUMANA en la base de datos. Se implementó un diseño gráfico que permite entrar con distintos usuarios, crearlos y administrarlos, además de eliminarse la necesidad de entrar a través de un archivo .csv. También se crearon validaciones y programación defensiva para evitar errores en caso se ingresen datos no válidos o espacios en blanco. En la Figura 1 se puede ver el modelo relacional que fue utilizado.

Actualmente, la mayor dificultad se encuentra en proveer de forma sencilla la aplicación de Matlab, que es la que actualmente se utiliza, a todo el mundo. Se necesita conectar con un repositorio en la nube para evitar posibles pérdidas, y facilitar el acceso en cualquier parte, ya que actualmente se puede lograr el acceso remoto pero desde las computadoras que poseen HUMANA. Además, el tener un los datos solo de manera local imposibilita que todos los usuarios compartan la misma información. También se necesita optimizar el almacenamiento de archivos ya que por ejemplo, el archivo edf consume demasiado espacio.

La red de cooperación latinoamericana de investigación, RedCLARA, ofrece un servicio de repositorios bajo la plataforma MiLab con el objetivo de apoyar en la preservación digital de datos de investigación y la colaboración en esta. El sistema específico para datos de investigación es el servicio "dataLab", creado en base al software "DataVerse". En este se pueden crear repositorios para el uso de la comunidad, de forma segura y de fácil acceso.[9].

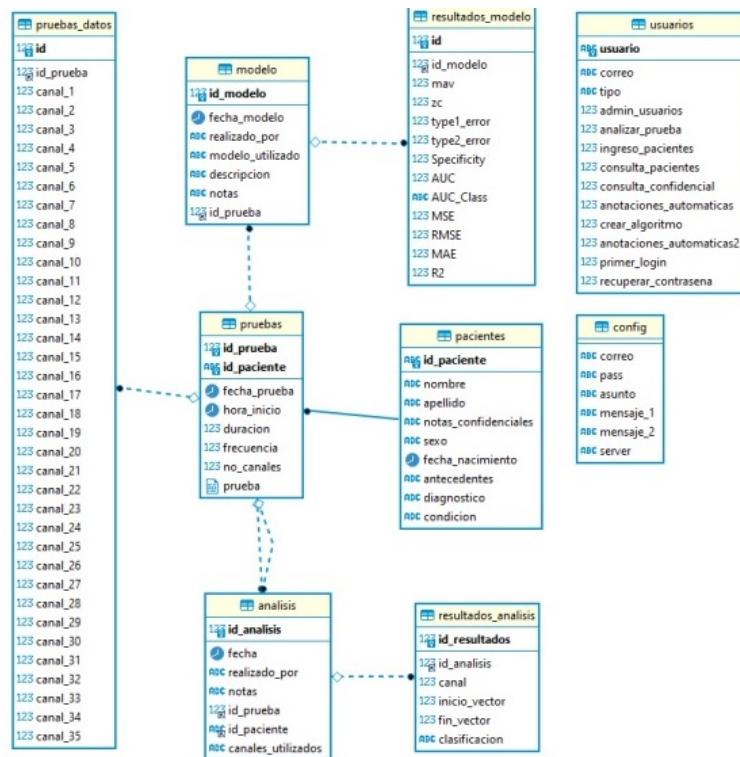


Figura 1: Modelo relacional utilizado en la fase anterior para la base de datos [7].



En las fases anteriores, sobre todo en la de Jorge Diego Manrique[7] se realizaron muchas mejoras en lo que respecta a la interfaz que permite el ingreso a la base de datos, tales como ampliar la cantidad de registros sobre el análisis, el paciente, las pruebas, el modelo utilizado, además de mejorar el ingreso de usuarios. Sin embargo, quedaron varias mejoras pendientes como mejorar la accesibilidad a la base de datos, optimizar la utilización de esta, además de expandir aún más su uso.

Actualmente la interfaz realizada por Jorge Diego Manrique está ajustada de buena manera a las herramientas de HUMANA, pero está pensada para un uso local, lo cual evita que se pueda compartir lo conseguido con la comunidad, además de limitar las posibilidades de esta aplicación.

Se busca crear un repositorio de señales y datos biomédicos relacionados a la epilepsia, de acceso público desde cualquier parte del mundo este debe estar bien organizado y con una interfaz sencilla para que cualquier persona pueda utilizar sus datos para sus propias investigaciones sobre la epilepsia.

El repositorio debe poder brindar a la comunidad aquello que ha investigado, y poder subir sus datos, con todos los datos que son necesarios tales como el método utilizado, el estado de lo tomado, el porcentaje de exactitud de los filtros en caso de tenerlo. De esta manera se busca aportar a la comunidad una herramienta que mejore el estudio de esta enfermedad de la que aún nos falta saber mucho.



#### **Objetivo General**

Desarrollar un repositorio de señales y datos biomédicos relacionados al estudio de la epilepsia, de acceso público en cualquier parte del mundo.

#### **Objetivos Específicos**

- Evaluar modelos de repositorios biomédicos existentes y determinar el más adecuado para la implementación del repositorio propuesto.
- Obtener señales y datos biomédicos de pacientes con epilepsia de HUMANA, sin incluir información privada o que pueda identificar a los pacientes.
- Determinar e implementar el procesamiento y la organización de los datos y los formatos en que se almacenarán.
- Implementar el repositorio y las herramientas para su manejo, administración, y la interacción con este.





Este proyecto consistió en mejorar lo realizado en la fase 2 por Jorge Diego Manrique, creando un repositorio que pueda ser de acceso al público desde cualquier parte del mundo, de manera que cualquiera pueda contribuir con nuevos datos, nuevos análisis, y además cualquiera pueda reproducir lo que se encuentra en nuestro repositorio.

Esto se logró a través de la herramienta Dataverse, a través del servicio que brinda la gente de Red Clara, con el cual se pudo crear un repositorio de carácter público, en el cual se puede controlar la forma en que la gente tiene acceso, a través del sistema de *dataverse* y *datasets* con lo cual se puede organizar de buena manera la información.

El proyecto cumple con implementar un repositorio en la nube, de carácter público, que sea accesible desde cualquier parte. La organización de este cumple con almacenar de forma ordenada, con su meta data correspondiente, la información sobre distintas pruebas y análisis de la epilepsia, a través de distintos métodos, siendo capaz de almacenar cualquier tipo de archivo. Las limitaciones son propias de Dataverse, el cual no permite hacer categorías propias para la meta data, esta solo puede seleccionarse. También el proceso de vinculación de *datasets* debe ser manual a través de una categoría en la meta data, lo cual limita un poco el enlazar varios *dataverse*.



### 6.1. Epilepsia

La epilepsia es una enfermedad que afecta al cerebro y las señales que envía. La perturbación en esta actividad neuronal causa en un aumento repentino de la actividad eléctrica que, puede causar convulsiones. Estas perturbaciones de la actividad neuronal varían según el área del cerebro donde ha ocurrido.

A pesar de ser una de las condiciones neurológicas más comunes, más precisamente, llegando a afectar a más de 50 millones de personas alrededor del mundo (siendo más común en niños y mayores de 60 años) no se ha logrado comprender esta condición en su totalidad, teniendo como prueba de esto que, a día de hoy, no se posee un tratamiento específico. De la población que sufre de epilepsia, un tercio sufren de epilepsia refractaria, lo cual se refiere a la clase de epilepsia en la que las convulsiones no pueden ser tratadas por los métodos antiepilépticos más utilizados o recomendados [1].

#### 6.1.1. Tipos de epilepsia

Existen más de 30 tipos de crisis epilépticas descritas con duración de unos segundos o minutos. Se pueden dividir en dos clases principales, convulsiones focales que afectan solamente una parte del cerebro, y convulsiones generalizadas que afectan todo el cerebro [2]. En la figura dos podemos ver un ejemplo de como se ve un encefalograma

En el caso de una crisis epiléptica focal o parcial, hay tres tipos: crisis parcial simple, compleja, y crisis parcial que desemboca en generalizada al extenderse a todo el cerebro. La crisis parcial simple es la que ocurre cuando se presenta alguna alteración de la memoria, el movimiento y las acciones. En este estado es probable que el afectado sepa lo que está pasando. La crisis parcial compleja es la más común, es en la que el afectado pierde el conocimiento, con repetición convulsiva de movimientos, es probable que el afectado no sepa que está sufriendo una convulsión. Normalmente, un médico puede interpretar que

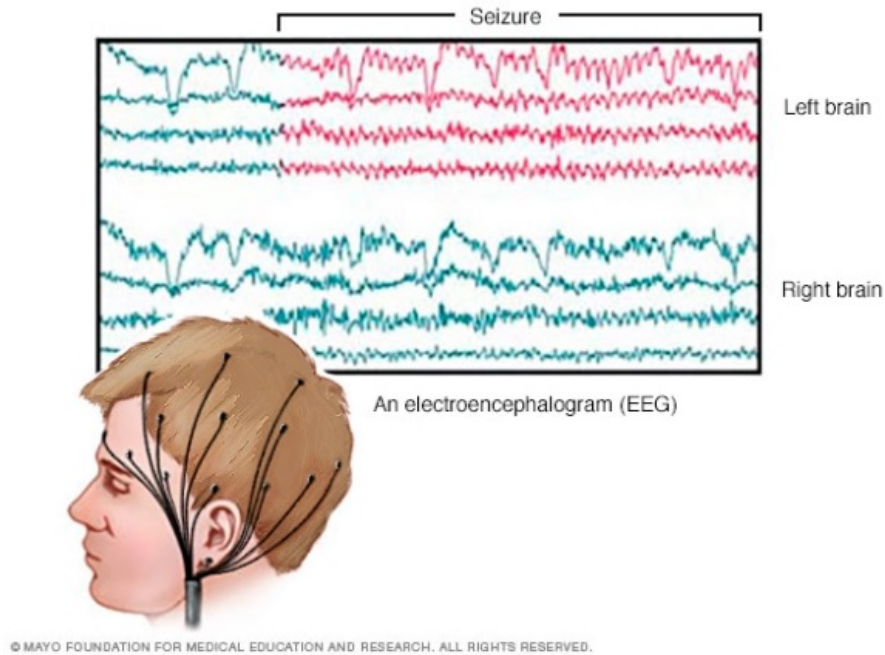


Figura 2: Actividad cerebral registrada por electroencefalograma [10].

clase de epilepsia sufre el paciente a través de un electroencefalograma (EEG) y generar un plan de tratamiento [2].

## 6.2. Electroencefalograma

Un electroencefalograma (EEG) es una prueba que detecta la actividad eléctrica del cerebro utilizando pequeños discos metálicos fijados en el cuero cabelludo. Las neuronas cerebrales se comunican a través de impulsos cerebrales y están activas en todo momento. Estas medidas se registran por medio de líneas onduladas en un reporte del encefalograma.

Un encefalograma puede ser utilizado para detectar cambios en la actividad cerebral, por medio de los cuales se pueden detectar enfermedades y trastornos cerebrales, especialmente la epilepsia u otros trastornos compulsivos. También puede detectar trastornos tales como daños cerebrales por lesiones en la cabeza, inflamación cerebral, trastorno de sueño, etc.

Un electroencefalograma también puede utilizarse para confirmar la muerte cerebral de una persona que se encuentra en un coma persistente. Además de esto, puede ser utilizado para encontrar la cantidad correcta de anestesia que se le debe dar a un paciente con coma inducido por medicamentos [10].

## 6.3. Base de datos

Cuando se habla de una base de datos relacional, se habla de que esta almacena y provee datos que están relacionados a otros. Se basan en un modelo fácil e intuitivo para acceder y almacenar la información en tablas. Cada fila es un registro con un número de identificación único conocido como llave y posee un atributo con información específica al registro [3].

El objetivo de una base de datos es recolectar la información y mantenerla de tal forma que pueda ser utilizada para análisis y operaciones futuras. Hay varios recursos que debe poseer como, por ejemplo, debe ser accesible para varios usuarios, la recolección debe ser conforme la información es generada, debe ser guardada la información aunque no sea usada en mucho tiempo, la lectura y escritura de información debe ser posible de hacer de forma constante, y la información debe estar relacionada para que sea fácil de buscar [3].

### 6.3.1. Conceptos básicos

*Schemas*: Es una colección de temas en una base de datos, tiene un dueño y puede contener diversos objetos dentro de la base de datos como tablas, vistas, secuencias, etc [11].

*Tablas*: Unidad básica de almacenamiento en la base de datos, está formada por columnas con diferentes tipos de datos, y filas con la información a almacenar [12].

*Tipos de datos*: tipo de información que se quiere almacenar en la columna indicada en la tabla, por ejemplo: INT (número entero), Float (número de punto flotante), String (campo alfanumérico), etc. [13]

*Llaves primarias*: combinación de columnas que identifican cada registro de la tabla para que pueda ser modificado o consultado fácilmente por separado o en conjunto por otros registros [14].

*Llaves foráneas*: estas relacionan las tablas entre sí. La llave foránea es llamada “hija” y define las columnas que son relacionadas en la tabla “padre” [15].

## 6.4. Base de datos en la nube

Una base de datos en la nube, es aquella que no se almacena en un equipo o sistema local, sino que se ejecuta desde la infraestructura de un proveedor de servicios.

Hay dos clases de bases de datos en la nube, está la clase tradicional que es similar a una base de datos *in-situ* y administrada internamente, con una pequeña diferencia en el origen de la infraestructura, ya que la compañía compra un espacio virtual de un proveedor de servicios en la nube. La organización se encarga de manejar la base de datos. Luego está la base de datos como servicio (DBaaS), donde el proveedor ofrece más servicios de gestión de base de datos [4].

### 6.4.1. Ventajas

Escalables: el proveedor del servicio puede aumentar los recursos o proporcionar más espacio de almacenamiento si así lo desea el cliente [4].

Integridad de la información: Este tipo de almacenamiento de información puede replicar instancias de la base de datos, lo cual permite que varios usuarios al mismo tiempo puedan usarla, sin que los datos pierdan integridad. [4].

Ahorro de espacio físico: no es necesario instalar un equipo dedicado local, ya que todo queda almacenado en los servidores del proveedor (de todas formas, tener un back-up local puede ser útil) [4].

Evitan problemas tales como la imposibilidad de acceder si se caen los servidores locales. Aumenta la seguridad: esto debido a que la información está protegida por empresas con reputación en el mundo digital [4].

Pero la ventaja más importante de todas es que se puede acceder a través de cualquier dispositivo desde cualquier parte del mundo [4].

## 6.5. Servicios

Se investigaron las mejores opciones para poder poseer el mejor servicio posible en cuanto a espacio, precio, y accesibilidad, entre ellos servicios como Microsoft Azure[16], Amazon s3[17] y en Oracle OCI[18]. En general los precios son similares, pero para Amazon se encontró mejor información sobre su uso, Además se encontró que el servicio de Amazon ya es utilizado para varias aplicaciones externas.

### 6.5.1. Dspace

*Dspace* es un software desarrollado por el instituto tecnológico de Massachusetts, el cual es desarrollado para organizar y manejar repositorios de ficheros, y es muy usado en las investigaciones de instituciones académicas. Esto se comprueba viendo que está siendo usado justamente por el repositorio de artículos de investigación de la universidad del valle de Guatemala. [19].

Se investigó como está hecho el repositorio de "physionetz" al revisarlo se descubrieron varias cosas, entre ellas, que los archivos médicos como los electroencefalogramas están guardados en formato edf (.european data format"), que es el más usado para temas de salud. También se descubrió que es desarrollada la página por el MIT, así como el software Dspace, teniendo muchas similitudes con este.

También se estudiaron los servicios que brinda LYRASIS, la compañía que gestiona Dspace, que tenía varios servicios de asistencia en el seleccionar un espacio para acoger el repositorio en la nube, sin embargo, los hemos considerado incompatibles con lo que estamos buscando por temas de precio y libertad en manejar el repositorio. [5].

Sin embargo, la mejor opción para el repositorio se encontró cuando se nos fue informado por medio del a Luis Roberto Furlan Collver de un proyecto llamado "Dataverse".

## 6.6. Dataverse

Dataverse es un proyecto de *open source*, desarrollado para guardar, archivar, manejar, citar y explorar data relacionada a la investigación. Esta desarrollada para facilitar el brindar tu información al resto del público interesado, además de facilitar el replicar el trabajo de otros usuarios [20].

El sistema con el que funciona consiste en que, Cada repositorio de Dataverse contiene múltiples archivos virtuales llamados colecciones de *dataverse*. Cada uno de estos *dataverse* contiene un conjunto de datos llamados *datasets*, que contienen archivos de datos Junto a los metadatos que el autor del repositorio considere necesarios para la reproducción de la información [20].

La perspectiva central de este proyecto fundado por Harvard con socios como Alfred P. Sloan Foundation, National Science Foundation, entre otros, es poder automatizar el trabajo de archivado, brindando crédito además de servicios de almacenamiento para los creadores de los datos. En el pasado, los investigadores tenían que escoger entre, mantener el crédito de su trabajo, pero tener que almacenar localmente por su cuenta toda la información, o contratar un servicio de repositorio que pudiera brindar su trabajo al público, pero perdiendo el crédito. Con esta herramienta, ese destino se evita [20].

### 6.6.1. DataLab

El servicio que estamos usando de forma específica es DataLab , desarrollado por Red-CLARA, la red de cooperación Latinoamericana en investigación. Está todavía en fase de desarrollo y nuestra participación en estos test pueden ayudar a que el proyecto mejore y obtenga características que aun no posee. Este servicio es parte de una plataforma llamada "MiLabçuyo objetivo es apoyar la gestión de datos, la preservación de la historia digital y el trabajo colaborativo de grupos de investigación asociados a RedCLARA. Datalab permite catalogar y conservar los datos en un lugar seguro y de fácil acceso, además de facilitar la cooperación grupal [9].

En la Figura 3 podemos ver el diagrama del servicio MiLab, y como se relacionan sus distintas secciones. Siendo DataLab la orientada a los datos de investigación, pero también teniendo otros sistemas de repositorio para Código(G-Lab) y para cálculo computacional (compLab) [9].

## 6.7. Aprendizaje Automático y la Epilepsia

El aprendizaje automático es altamente utilizado en las áreas de la salud y la biología, ya que permite el análisis de mucha información. Dependiendo del modelo utilizado y la

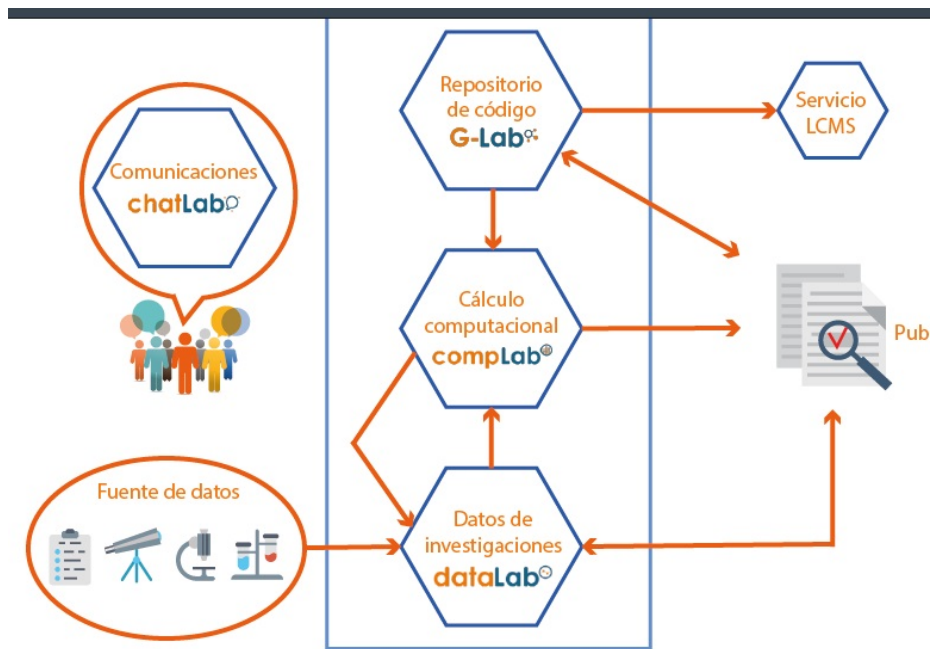


Figura 3: Servicios de Milab [9].

aplicación que se le dé a este modelo, pueden llegar a responderse una gran cantidad de cuestiones. Un modelo de clasificación, por ejemplo, permite por medio de datos históricos, determinar si algo va a suceder o no. Un modelo de series de tiempo permite predecir valores de una variable que depende del tiempo tomando en cuenta parámetros como la tendencia, ciclos, etc. Un modelo de *clustering* permite agrupar en base a características. Hay una gran variedad de modelos y aplicaciones [21].

Este método se puede utilizar para la clasificación y detección de epilepsia. También se ha utilizado para análisis de señales EEG ya que de esta forma se puede descubrir información importante de la señal [1].



---

### Análisis Servicios:

---

Antes de Realizar un prototipo habiendo escogido un modelo, se realizó un análisis de los distintos sistemas de repositorio que podíamos utilizar, para realizar el repositorio público de epilepsia. Lo principal que se busca es que, el servicio con el que almacenemos nuestros datos en la nube nos permita darle acceso al almacenamiento a distintos actores, de manera fácil y sencilla. El precio debe ser razonable y debe permitirnos almacenar los archivos edf que queremos, teniendo en cuenta el peso que estos archivos pueden tener. Para el almacenamiento de datos, analizamos 4 servicios, los de Amazon, Microsoft, y Oracle.

#### 7.1. Amazon AWS

De entre los servicios investigados y analizados, este es el mejor. Los servicios de Amazon permiten mayor versatilidad, teniendo un gran catalogo entre sistemas de base de datos SQL y NoSQL. Una de las opciones es Aurora, el DBMS relacional de propósito general de Amazon, se recupera automáticamente y puede llegar a almacenar hasta 64 TB. Esta plataforma incluso es capa de integrar aprendizaje automático a través de SQL. Tiene un coste de 0.10 dólares por GB de almacenamiento, 0.20 dólares por operaciones de entrada y salida, y en general, el coste es razonable. También, las ofertas de migración de base de datos de Amazon son las mejor es en el mercado. El sistema de migración de AWS es una utilidad de replicación integral, que le permite inicializar su plataforma DBaaS con datos y luego mantenerla sincronizada con su sistema fuente. También permite configurar la replicación con una gran variedad de plataformas heterogéneas. Amazon incluso provee una herramienta que convierte el esquema de origen de la base de datos, con su contraparte de Amazon [17].

## 7.2. Microsoft Azure

Microsoft incluye entre su plataforma de base de datos, la base de datos SQL de Azure y el almacén de datos SQL para aplicaciones de big data. Actualmente la capacidad de almacenamiento de Azure es de 100 TB en la base de datos Azure SQL. Esta tiene precios de por ejemplo 0.12 dólares por GB al mes en base de datos con redundancia local, y con redundancia zonal 0.23 dólares. Considerando también las opciones NoSQL que ha ido agregando Azure, en general Microsoft Azure tiene un mayor precio [22].

## 7.3. Oracle

El enfoque principal de Oracle es promover sus plataformas de base de datos autónomas y el almacén de datos autónomo. Oracle ofrece que su propuesta puede realizar gran parte de la configuración de la base de datos, el ajuste, la aplicación de parches y el trabajo de actualización que generalmente realizan los administradores de la base de datos. Incluye la mayoría de las ofertas que hemos visto anteriormente en los otros dos servicios. En donde falla Oracle, es en que ofrece pocas facilidades y herramientas para la migración de base de datos. En general, ofrece poca flexibilidad en incluirse servicios ajenos a Oracle con servicios de esta misma [18].

## 7.4. introducción a DataLab

Todo este análisis descrito anteriormente, fue previo a la introducción a Datalab de parte de la gente de RedClara , un ambiente creado dentro de la plataforma Dataverse, en el cual ellos nos brindan el servicio de almacenamiento , con el cual interactuamos a través de una interfaz que permite por medio de los llamados "Datasetz "Dataverse", desarrollar un repositorio con la meta data adecuada. Antes de esto, por supuesto, se realizaron pruebas de manera local para diseñar la forma en que se realizaría dentro de dataverse el repositorio .

Antes de desarrollar el prototipo del que se va a hablar, se realizaron multiples pruebas dentro del software MySQL Workbench 8.0 con el fin de probar las diferentes formas en que se pueden relacionar tablas en una base de datos. Estas pruebas sirvieron para definir el sistema de tres tablas que fue utilizado en el prototipo final, argumentando tres categorías principales para el repositorio, Una siendo Usuarios, ya que se debe mantener registro de quien usa el repositorio. Otra categoría es Pruebas y la ultima Análisis. Estas dos se decidieron definir por separado, argumentando que , como se pudo observar en los sistemas de las fases anteriores de este proyecto, un Análisis de por ejemplo, aprendizaje automático, puede usar datos recopilados por múltiples usuarios.

El primer prototipo desarrollado se realizó para comprobar de forma local, la manera en que pueden ser subidos los archivos de por ejemplo electroencefalogramas, a una base de datos. Esto se comprobó debido a que pueden llegar a ser muy pesados los análisis de epilepsia, sobre todo cuando engloban un análisis de varias etapas, agrupando múltiples archivos edf.

Para esto primero se creó una tabla sencilla con Microsoft SQL Mangement Studio, como se puede ver en la figura 4, con el único objetivo de poder guardar en una tabla, los archivos que se suben con la aplicación de Visual Studio.

El siguiente paso fue el propio desarrollo de la app, utilizando las herramientas de Visual Studio. El código consiste varios botones con los cuales, se puede navegar a través de los archivos locales del PC, y se puede activar una rutina de conexión SQL, para guardar los archivos en la base de datos. Se utilizo una interfaz como la que se ve en la Figura 5.

Lo que pudimos comprobar usando este sistema, es que el peso de los archivos puede llegar a causar varios problemas en el tiempo de guardado. Sin embargo, esto puede ser reducido a partir de comprimir los archivos ,en un archivo zip, el cual no causa conflicto.

Column Name	Data Type	Allow Nulls
id	int	<input type="checkbox"/>
Data	varbinary(MAX)	<input checked="" type="checkbox"/>
Extencion	char(4)	<input checked="" type="checkbox"/>
		<input type="checkbox"/>

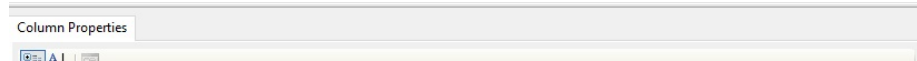


Figura 4: tabla para pruebas.

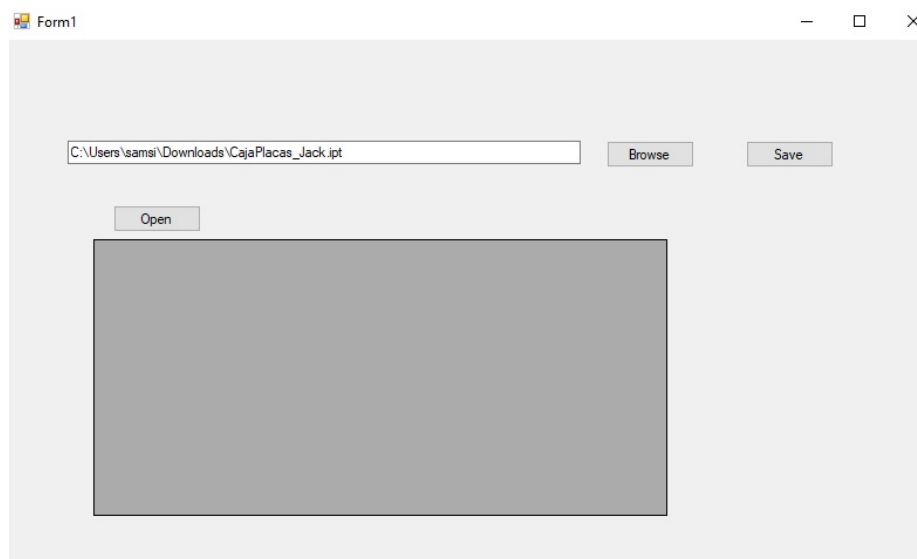


Figura 5: app ddesarrollada en visual studio.

Gracias a esta prueba, se pudo comprobar la necesidad de definir un protocolo para el ingreso de archivos al repositorio final. En este caso se ha decidido que todos los archivos que se puedan agruparen fases, de manera que no causen conflicto con organizar la meta data, deberán ser comprimidos en un archivo zip.

Prototipo 2: Dataverse

---

En este capítulo se presenta el prototipo realizado ya en el software que se decidió utilizar. Se buscó mantener una estructura parecida a lo realizado en etapas anteriores de forma local. Un detalle importante para considerar es que este servicio ya incluye secciones de meta data, que están dadas a ser incluidas de forma opcional, pero no es posible agregar una sección personalizada, por lo que se tuvo que adaptar el diseño a las categorías de meta data ya establecidas.

Para realizar el repositorio, Primero hemos realizado un *dataverse* general, en el cual no van a ser colocados *datasets*, sino que va a estar separado por tres *dataverse*, uno en la que cada usuario tendrá un perfil, donde podrá ir adjuntando todos los *datasets* que vaya realizando en los demás data verse, un dataverse para Pruebas, es decir, los datos en bruto de

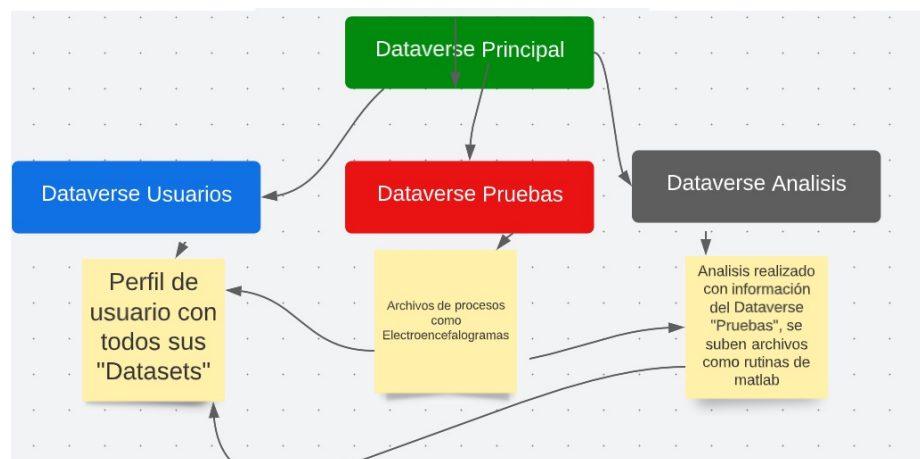


Figura 6: Relación entre los *Dataverse* del repositorio.

Electroencefalogramas y otros métodos para analizar los síntomas de la epilepsia, y otro para Análisis, ya que como podemos saber, un conjunto de datos puede ser pasado por múltiples métodos de análisis como por ejemplo, distintas rutinas de Aprendizaje Automático. Estas relaciones las podemos ver en el diagrama de la Figura 6

## 9.1. Dataverse: Pruebas

Este *dataverse* fue diseñado para ser el centro de todos los archivos, y ser adjuntado en otros *dataverse*. Un ejemplo de un *dataset* creado para esta sección se encuentra en la Figura 7. La Meta dada que requiere cada dataset de este dataverse es la siguiente:

*Id del dataset*: Id con el que se identifica el *dataset*.

*título*: debe ser un título fácil de buscar.

*Autor*: Aquí va el autor del dataset, no el autor del archivo que se está subiendo, considerando que un grupo de archivos edf, se incluye también un id de usuario.

*contact*: Correo del autor.

*Description*: Una descripción general que quiera el autor sobre el archivo o archivos.

*Subject*: Esta categoría es obligatoria en Dataverse, considerando el estudio de la epilepsia, se debe colocar “Medicine, Health and Life Sciences

*Keyword*: En keyword se debe colocar el número de capas de electroencefalograma en caso de ser un encefalograma en archivo edf

*idioma*: Idioma en que fue realizado todo el proceso.

*Producer*: Persona que realizó la toma de datos.

*Producción date*: Fecha en que se obtuvieron los datos.

*Producción place*: Lugar donde fueron obtenidos estos datos.

*Contributor*: Paciente del que fueron obtenidos los datos.

*Deposit date*: Día en el que fue realizada la entrada del dataset.

*Time Period Covered*: Periodo de tiempo estudiado.

*Kind of data*: Tipo de datos tomados para el análisis y estudio de la epilepsia, por ejemplo, si el archivo es un edf de un electroencefalograma, se coloca “electroencefalograma”.

*Origin of Sources*: Información de la fuente del archivo, en este caso, del paciente.

*Data source*: En caso de estar ingresando al repositorio un archivo externo, en esta sección iría el link de por ejemplo, un archivo de physionet.



Dataset Persistent ID 	doi:10.21348/FK2/C3UPMC
Title 	Analisis de Pablo
Author 	Silvestre, Samuel (student uvg) - ResearcherID: 03439349
Contact 	Use email button above to contact. Silvestre, Samuel (uvg)
Description 	*descripcion del analisis (2022-08-30)
Subject 	Computer and Information Science; Medicine, Health and Life Sciences
Notes 	*descripcion del algoritmo*
Language 	English
Production Date 	2022-10-29
Contributor 	Researcher : David Vela
Depositor 	Silvestre, Samuel
Deposit Date 	2022-09-08
Time Period Covered 	Start: 2022-09-05 ; End: 2022-09-07
Date of Collection 	Start: 2022-08-23 ; End: 2022-08-26
Kind of Data 	machine learning no supervisado
Software 	matlab, Version: 2022
Related Datasets 	<a href="https://dataverse.redclara.net/dataset.xhtml?persistentId=doi:10.21348/FK2/DEVK1C&amp;version=DRAFT">https://dataverse.redclara.net/dataset.xhtml?persistentId=doi:10.21348/FK2/DEVK1C&amp;version=DRAFT</a>

Figura 7: Ejemplo dataset de edf de un paciente.

## 9.2. Dataverse: Análisis

En este *dataverse*, es donde se ingresa todo análisis realizado con base en los archivos ingresados en el *dataverse* de Pruebas. Esto puede incluir por ejemplo métodos de aprendizaje automático, o métodos más simples, que utilicen archivos incluidos en el *dataverse* de pruebas. Un ejemplo de un *dataset* creado para esta sección se encuentra en la Figura 8

*Id del dataset*: Id con el que se identifica el *dataset*.

*título*: Debe ser un título fácil de buscar. *Autor* Aquí va el autor del dataset, no el autor de la rutina del análisis , se incluye también un id de usuario.

*contact*: Correo del autor. *Description*: Descripción del análisis y estudio realizado que se está ingresando a nuestro repositorio.

*Subject*: Esta categoría es obligatoria en Dataverse, considerando el estudio de la epilepsia, se debe colocar “Medicine, Health and Life Sciences.

*Notes*: Datos importantes del algoritmo, por ejemplo, márgenes de error.

*language*: Idioma en que fue realizado el proceso.

*production date*: Día en que fue finalizado el análisis.

*contributor*: Autor del análisis.

*Deposit date*: Fecha en que es depositado el análisis en nuestro repositorio.

*Time Period Covered*: Tiempo que tomó en ser realizado el análisis.

*Kind of Data* : Tipo de análisis realizado, por ejemplo, aprendizaje automático no supervisado.

<b>Dataset Persistent ID</b> ?	doi:10.21348/FK2/DEVK1C
<b>Title</b> ?	EEG de Pablo
<b>Author</b> ?	Silvestre, Samuel (student uvg)
<b>Contact</b> ?	Use email button above to contact. Silvestre, Samuel (uvg)
<b>Description</b> ?	kjfdkkek (2022-08-30)
<b>Subject</b> ?	Medicine, Health and Life Sciences
<b>Keyword</b> ?	5
<b>Language</b> ?	English
<b>Producer</b> ?	Silvestre, Samuel (uvg)
<b>Production Date</b> ?	2022-08-24
<b>Production Place</b> ?	uvg
<b>Contributor</b> ?	Related Person : Jaime sandoval
<b>Depositor</b> ?	Silvestre, Samuel
<b>Deposit Date</b> ?	2022-09-08
<b>Time Period Covered</b> ?	Start: 2022-08-15 ; End: 2022-08-17
<b>Kind of Data</b> ?	electroencefalograma

Figura 8: Ejemplo dataset Analisis de un edf.

*Software:* Software utilizado en el análisis, como por ejemplo Matlab.

*Related Datasets:* Link directo a los *datasets* que contienen los archivos utilizados en el análisis.

### 9.3. Dataverse: Usuario

En esta *dataverse*, para poder acceder a utilizar nuestro repositorio, deberán crear un dataset de su usuario, incluyendo en la meta data, información básicas (nombre, id, contacto y una descripción sobre la labor del usuario) se incluye la sección de related publications, en caso de haber usado un link externo para por ejemplo, un análisis, y la sección de related datasets, donde deberán ir adjuntando todos los datasets que creen dentro del repositorio. Un ejemplo de una entrada en este *datverse* se puede ver en la Figura 9.



dataLab > UVG > Repositorio Epilepsia 1.0 > Usuarios >

## Samuel Silvestre

**Draft** **Unpublished**



Silvestre, Samuel, 2022, "Samuel Silvestre", <https://doi.org/10.21348/FK2/U3ABRX>, dataLab, DRAFT VERSION

Cite Dataset ▾

[Learn about Data Citation Standards.](#)

Publish Dataset	
Edit Dataset ▾	
Contact Owner	Share

Dataset Metrics ⓘ

0 Downloads ⓘ

**Description** ⓘ

ssadasdsadsad (2022-06-30)

**Subject** ⓘ

Medicine, Health and Life Sciences

**Keyword** ⓘ

po

**Related Publication** ⓘ

<https://dataverse.redclara.net/dataset.xhtml?persistentId=doi%3A10.21348%2FFK2%2FN3QVEL&version=DRAFT> url: 1

**Notes** ⓘ

sdasdas

Figura 9: Ejemplo perfil de usuario.



---

### Prototipo 3: Nuevo Dataverse

---

Para este prototipo, lo que se realizó fue que se realizó en un *dataverse* en el que se nos fue entregado el poder máximo para poder modificar apariencia, agregar widgets, y probar funciones que no pueden ser probadas sin publicar un *dataverse*. También se modificó la estructura del *dataverse* de “análisis” basándonos en lo realizado en la tesis, en desarrollo al momento de realizar este trabajo, por Camila Lemus. Es decir, está hecha para análisis de machine learning en general. Un ejemplo de un *dataset* creado para esta sección se encuentra en la Figura 10

*Title*: un título que sea fácil de buscar.

*Autor*: Nombre del autor, afiliación, id y clase de id.

*Contact*: información de contacto del autor, incluyendo ID y correo electrónico.

*Description*: datos generales del análisis, tales como que fue utilizada una exactitud RNA, luego estos datos pueden ser incluidos en palabras clave para facilitar su búsqueda.

*Subject*: siempre debe ser “Medicine, health and life science”.

*Keyword*: es obligatorio agregar la cantidad de razones y cuáles fueron las razones utilizadas en el análisis.

*Production date*: fecha en que fue realizado el análisis.

*Production place*: donde fue realizado.

*Kind of data*: El tipo de datos utilizados, por ejemplo, electroencefalogramas.

*Series*: Es la categoría que se ha escogido para incluir el porcentaje de exactitud, en notas, de manera opcional, se puede colocar el tiempo de procesamiento.

<b>Dataset Persistent ID</b> ?	doi:10.21348/FK2/NRZPYI
<b>Title</b> ?	Analisis_juan_1
<b>Author</b> ?	Silvestre, Samuel (student uvg) - ResearcherID: 03439349
<b>Contact</b> ?	Use email button above to contact. Silvestre, Samuel (uvg)
<b>Description</b> ?	Exactitud RNA, Ubonn - Sano (2022-08-30)
<b>Subject</b> ?	Medicine, Health and Life Sciences
<b>Keyword</b> ?	3 (razones1,2, y 3)
<b>Production Date</b> ?	2022-08-24
<b>Production Place</b> ?	uvg
<b>Depositor</b> ?	Silvestre, Samuel
<b>Deposit Date</b> ?	2022-10-25
<b>Kind of Data</b> ?	electroencefalogramas
<b>Series</b> ?	99.99%: tiempo de 8.32
<b>Software</b> ?	Matlab, Version: 2021
<b>Related Datasets</b> ?	<a href="https://dataverse.redclara.net/dataset.xhtml?persistentId=doi:10.21348/FK2/DEVK1C&amp;version=DRAFT">https://dataverse.redclara.net/dataset.xhtml?persistentId=doi:10.21348/FK2/DEVK1C&amp;version=DRAFT</a>

Figura 10: Ejemplo Nuevo formato de Dataset de Análisis

*Software:* El software utilizado para el análisis, como por ejemplo, Matlab.

*Related datasets:* Datasets relacionados a este análisis, es decir, que se hayan utilizado, como, por ejemplo, archivos edf que se encuentren en el dataset de pruebas.

---

## Nuevo dataverse de Pruebas

---

Tuvimos una reunión con de HUMANA donde les pedimos consejo sobre que considerarían que le haría falta a la metadata de las pruebas de epilepsia, es decir al Dataverse de Pruebas. Luego de recibir la retroalimentación, esta fue la nueva estructura de la metadata en esta sección del repositorio. Un ejemplo de un *dataset* creado para esta sección se encuentra en la Figura 11

*título:* Debe ser representativo del archivo que se está subiendo, de manera que sea fácil de buscar.

*Author:* La información de autor debe incluir el nombre, institución a la que está relacionado, y el id que se le entregó en el repositorio.

*Contact:* Información de contacto de la misma forma.

*Description:* La descripción debe ser corta y concisa. Debe contener información relevante sobre la manera en que se produjo la muestra.

*Subject:* El “subject” siempre debe ser el que está en la imagen.

*Keyword:* En keyword incluimos los datos más concretos que faciliten la búsqueda, como la frecuencia en que el electroencefalograma fue tomado, el canal en que fue hecho, y la cantidad de tomas hechas. Esto con el objetivo de que sea más fácil encontrar cierta clase de datos.

*Notes:* En notes, se pone el tiempo que requirió adquirir esta información.

*Language,* el idioma en que fue hecho.

*Contributor:* El origen de los datos, como por ejemplo physionet en este caso, o HUMANA en otros.















Dataset Persistent ID 	doi:10.21348/FK2/ZNUSZH
Title 	PN14 Siena Scalp EEG Database
Author 	Silvestre, Samuel (student uvg) 03439349
Contact 	Use email button above to contact. Silvestre, Samuel (uvg)
Description 	Male patient , 4 WIAS seizures, left lateralization (2022-08-30)
Subject 	Medicine, Health and Life Sciences
Keyword 	512 HZ (eeg channel 29) 4 tomas
Notes 	1408 minutes
Language 	English
Contributor 	Hosting Institution : Physionet
Depositor 	Silvestre, Samuel
Deposit Date 	2022-11-08
Kind of Data 	ELECTROENCEFALOGRAMA
Characteristic of Sources Noted 	Seizure n 1 : File name: PN14-1.edf Registration start time: 11.44.58 Registration end time: 13.56.06 Seizure start time: 13.46.00 Seizure end time: 13.46.27 Seizure n 2 : File name: PN14-2.edf Registration start time: 15.50.13 Registration end time: 18.22.58 Seizure start time: 17.54.52 Seizure end time: 17.55.04 Seizure n 3 : File name: PN14-3.edf Registration start time: 16.17.45

Figura 11: Ejemplo nuevo Dataset de Pruebas

*Depositor*: El que está subiendo el dataset .

*Deposit date*: Día en que se crea el dataset.

*Kind of data*: El tipo de archivo que se está subiendo, por ejemplo electroencefalograma.

*Notas*: Información general del tiempo de cada toma.

---

### Cómo Utilizar el Dataverse

---

Debido a las posibles dificultades que podrían existir a la hora de ingresar a este repositorio, y utilizar todas sus virtudes, se realizó una guía que detalla paso a paso el uso de este repositorio. Esta guía no será expuesta en su totalidad en este documento, ya que sería redundante con información presentada anteriormente.

Como primer paso, debido a que este repositorio se realizó en la plataforma de *Datalab*, se debe crear una cuenta en esta. Para realizar una cuenta se tienen que seguir los pasos tradicionales de cualquier página, es decir, llenar requisitos de Usuario, contraseña, información como el nombre, correo y afiliación del nuevo usuario, etc. Es fundamental realizar esto, o el repositorio no podrá ser utilizado. La sección de creación de usuario se encuentra en la figura 12.

Luego, para obtener acceso al *dataverse*, no hay que realizar ningún paso, se llegó a la conclusión que requerirle al usuario que mandara un correo pidiendo acceso, podría llegar a ralentizar el proceso del repositorio, volviendolo poco eficiente en uno de sus objetivos principales, que es utilizarlo para mejorar la investigación sobre la epilepsia en la comunidad. Sin embargo, es recomendable observar el comportamiento de los usuarios cuando este repositorio sea publicado, ya que si no se mantiene un orden, tendrá que aplicarse esto.

Cada usuario deberá crearse un *dataset* de usuario en el dataverse designado, en el cual solo incluirá su nombre, contacto, y los links de los *datasets* que haya creado, para facilitar la búsqueda de por ejemplo, un trabajo en específico de cada usuario. Este requisito no será obligatorio. Un ejemplo de esto está en la figura 13.

Para ingresar un archivo de electroencefalograma por ejemplo, o un archivo de Matlab de *Deep learning*, se debe entrar al *dataverse*, e ingresar el archivo con toda la metadata requerida. De considerar necesario agregar más información de la requerida, se puede utilizar campos de metadata que no son obligatorios, luego de haber creado el *dataset*.

**Username \*** ⓘ Create a valid username of 2 to 60 characters in length containing letters (a-z), numbers (0-9), dashes (-), underscores (\_), and periods (.).

**Password \*** ⓘ Your password must contain:

- At least 6 characters (passwords of at least 20 characters are exempt from all other requirements)
- At least 1 character from each of the following types: letter, numeral

**Retype Password \*** ⓘ

**Given Name \*** ⓘ

**Family Name \*** ⓘ

**Email \*** ⓘ

**Affiliation \*** ⓘ


**Position \*** ⓘ

**General Terms of Use \*** ⓘ There are no Terms of Use for this Dataverse installation.

Figura 12: Página donde se crea una cuenta.

## Samuel Silvestre

**Draft** **Unpublished**



Silvestre, Samuel, 2022, "Samuel Silvestre", <https://doi.org/10.21348/FK2/5AJFPG>, dataLab, DRAFT VERSION

ⓘ

Cite Dataset ▾ [Learn about Data Citation Standards.](#)

**Description** ⓘ \*estudiante mecatronica (2022-11-26)

**Subject** ⓘ Medicine, Health and Life Sciences

Files Metadata Terms Versions

Figura 13: Página donde se crea una cuenta.



Figura 14: Ejemplo de como se edita un dataset.



Figura 15: Archivo pdf subido al repositorio

Cualquier problema con los *datasets* creados, La plataforma cuenta con un botón de contacto, con el cual, se puede consultar via correo con el dueño del *dataverse* o el *dataset*, por ellos es muy importante que se agregue correo de contacto en cualquier ingreso de archivos.

La razón por la que , como se menciona en artículos anteriores,, se requiere cierta información en forma de *keywords*, es para facilitar la búsqueda de una clase especifica de archivos, ya que el repositorio cuenta con un buscador, el cual se puede usar de forma que encuentre una palabra en cualquier *dataset*, o que se utilice la búsqueda avanzada, que en este caso, es buscar una palabra como por ejemplo "17 canales.<sup>en</sup> una categoría de metadata específica, como por ejemplo *keywords*, esto se puede ver mejor en la figura 14.

Cabe destacar el formato en que se presentan los distintos tipos de archivos que se pueden subir al repositorio , cuando se sube un archivo edf o archivos más pesados, se recomienda utilizar compresión rar para reducir el tiempo de carga, por ejemplo como en la figura 15. Cuando se sube un archivo pdf se ve como en la figura 16. .Para subir imágenes, se debe usar formato rar para comprimir las y subirlas, o subirlas como pdf, ya que la plataforma permite imágenes pero solamente para el icono que aparezca al buscar el repositorio.

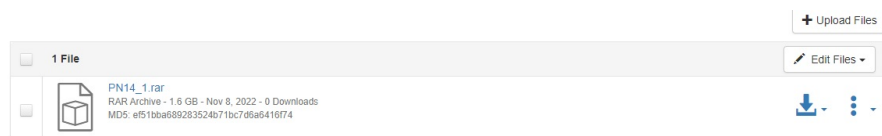


Figura 16: Archivo rar subido al repositorio

## CAPÍTULO 13

---

### Conclusiones

---

- Se logró desarrollar un repositorio de señales y datos biomédicos relacionados al estudio de la epilepsia, de acceso público en cualquier parte del mundo.
- Se evaluaron la mayor cantidad posible de las posibilidades para elaborar el repositorio y se consiguió la mejor opción
- se logró incluir señales y datos biomédicos de pacientes con epilepsia de HUMANA, sin incluir información privada o que pueda identificar a los pacientes.
- Se logró mantener de forma segura el repositorio de la mano de la gente de DataLab
- Se logró planear de manera correcta y eficiente organización de los datos
- se logró Implementar el repositorio y las herramientas para su manejo, administración, y la interacción con este.



- Se recomienda platicar con la gente de RedClara sobre la posibilidad de crear secciones de Metadata personalizadas para facilitar aún más el ingreso de esta, esto sería de mucha utilidad, ya que en la situación actual, se tuvo que adaptar muchas secciones de metadata a lo que necesitamos, y teniendo en cuenta los términos, a los usuarios se les tendrá que explicar qué significa cada cosa.
- Se recomienda conversar con el proveedor maneras de automatizar la clasificación de la data entre los *dataverse*, esto para que los usuarios puedan seguir menos pasos a la hora de utilizar nuestro repositorio.
- Se recomienda mejorar el diseño estético del repositorio , esto no es una prioridad, pero si la interfaz es más agradable a la vista, más sencillo será utilizarla.
- Se recomienda evitar un exceso de requerimientos a la hora de ingresar un archivo, sobre todo no redundar en categorías, ya que si se cae en eso, solamente se está alargando el proceso para el usuario de forma innecesaria.
- Se recomienda llevar un seguimiento de la vida del repositorio , y el desempeño de las reglas actuales de este, para poder ajustar el proceso para ingresar en caso de ser necesario.



- [1] C. E. Stafstrom, “Seizures and epilepsy: an overview for neuroscientists,” *PubMed*, 2015.
- [2] BUPA, “Epilepsia,” *BUPA Global Latinoamerica*, 2021.
- [3] Oracle, “What is a Relational Database?” <https://www.oracle.com/database/what-is-a-relational-database/>, 2022.
- [4] ———, “Base de datos en la nube,” <https://www.oracle.com/es/database/what-is-a-cloud-database/>, 2022.
- [5] Physionet, “<https://physionet.org/about/>,” 2022.
- [6] HUMANA, “Especialistas en enfermedades neurológicas de difícil control,” *humanagt.org.*, 2022.
- [7] J. Manrique, “Herramienta de Software con una Base de Datos Integrada para el Estudio de la Epilepsia - Fase II,” Tesis de licenciatura, Universidad Del Valle de Guatemala, 2021.
- [8] M. Pineda, “Diseño e Implementación de una Base de Datos de Señales Biomédicas de Pacientes con Epilepsia,” Tesis de licenciatura, Universidad Del Valle de Guatemala, 2021.
- [9] RedCLARA, “MiLab,” <https://redclara.net/index.php/es/servicios-rc/milab>, 2022.
- [10] M. Clinic, “EEG (electroencefalograma),” <https://www.mayoclinic.org/es-es/tests-procedures/eeg/about/pac-20393875/>, 2022.
- [11] Oracle, “Introduction to Schema Objects,” <https://docs.oracle.com/en/database/oracle/oracle-database/19/admqs/managing-schema-objects.htmlGUID-8AC1A325-3542-48A0-9B0E-180D633A5BD1>, 2022.
- [12] ———, “Managing Tables,” <https://docs.oracle.com/cd/B1930601/server.102/b14231/tables.htm>, 2022.
- [13] ———, “Selecting a Datatype,” <https://docs.oracle.com/cd/A5861701/server.804/a58241/ch5.htm>, 2022.

- [14] —, “Primary Keys,” <https://docs.oracle.com/en/database/other-databases/nosql-database/12.2.4.5/java-driver-table/primary-keys.html> `GUID-6A063474-4A3A-4981-B1D8-71D0D95BE8DF`, 2022.
- [15] —, “[https://docs.oracle.com/cd/E17952\\_01/mysql-5.6-en/create-table-foreign-keys.html](https://docs.oracle.com/cd/E17952_01/mysql-5.6-en/create-table-foreign-keys.html),” 2022.
- [16] —, “Calculadora de precios,” <https://www.oracle.com/es/cloud/pricing/>, 2022.
- [17] Amazon, “Precios de AWS,” <https://aws.amazon.com/es/pricing/?nc2=hqlprln>, 2022.
- [18] Oracle, “technology pricing list,” <https://www.oracle.com/assets/technology-price-list-070617.pdf>, 2022.
- [19] DSpace, “<https://physionet.org/about/>,” 2022.
- [20] Dataverse, “<https://dataverse.org/about>,” 2022.
- [21] M. K. Siddiqui, “A review of epileptic seizure detection using machine learning classifiers,” *PubMed*, vol. 7, n.o 1, pág. 5, 2020.
- [22] Microsoft, “Precios de Azure,” <https://azure.microsoft.com/es-es/pricing/product-pricing>, 2022.



### 16.1. Repositorio Dataverse

En el siguiente enlace se puede acceder al repositorio creado en este trabajo.

<https://dataverse.redclara.net/dataverse/epil10>

### 16.2. Repositorio Github

Este es el repositorio de Github donde se incluyen los archivos utilizados en las etapas de este proyecto para su elaboración

<https://github.com/Samuelsil997/Repositorio-Biomedico-2022>